DOCUMENT RESUME

ED 358 158 TM 019 957

AUTHOR Rule, David L.
TITLE A Simulation-Based Comparison of Several Stochastic Linear Regression Methods in the Presence of Outliers.
PUB DATE Apr 93
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; *Bayesian Statistics; Comparative Analysis; *Computer Simulation; *Estimation (Mathematics); Goodness of Fit; Least Squares Statistics; *Mathematical Models; Matrices; *Regression (Statistics); Research Methodology; Sample Size; Scores
IDENTIFIERS Bootstrap Methods; *Outliers; *Weighted Structural Regression

ABSTRACT

Several regression methods were examined within the framework of weighted structural regression (WSR), comparing their regression weight stability and score estimation accuracy in the presence of outlier contamination. The methods compared are: (1) ordinary least squares; (2) WSR ridge regression; (3) minimum risk regression; (4) minimum risk 2; (5) goodness of fit index (GFI); and (6) WSR reduced rank regression. Three population covariance matrices were used that were drawn from applied behavioral science literature as the basis for generating samples, some of which were contaminated. A bootstrap method was used to compare the regression methods. Analysis resulted in 4 sets of bootstrap samples for each of the population systems, 12 in all. Results support the notion of increased efficacy of the adaptive forms of WSR in small sample applications where outlier contamination exists. The improvement over conventional least squares is not always substantial, but it is notable that adaptive forms of WSR based on the concept of empirical Bayes covariance estimation can provide consistent and sometimes substantial improvement over conventional methods. Five tables present analysis data. Appendix A provides the basis for the WSR class of methods. Appendices B and C each contain three tables of analysis information. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED358158

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
DAVID L. RULE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Simulation-Based Comparison of Several Stochastic Linear Regression Methods in the Presence of Outliers

David L. Rule

Departments of Educational
Psychology, Psychology, and Teacher Education
Marist College, Poughkeepsie, NY

Paper Presented at the Annual Conference of the
American Educational Research Association
April 16, 1993

TM019951

Introduction

One of the major classes of statistical methods for dealing with relationships between variables in a prediction context is multiple regression. A difficulty that researchers often encounter when they initially decide to use multiple regression is the potentially large number of competing models or approaches for using the methods. For example, possible regression models might include least squares and its many variations (e.g., forward, backwards, all subsets), ridge regression, reduced rank regression, or one or more of many structural model approaches (e.g., LISREL or EQS). Additionally, least squares (OLS), perhaps the most common of these regression methods, as evidenced by major texts in the area (e.g., Darlington, 1990; Dillon & Goldstein, 1984; Pedhazur, 1982), has as an underlying tacit assumption the observed variables contatin no measurement error.

One common type of variable contamination that the researcher using OLS regression -- or any regression method -- is acquainted with is the outlier. The issue of what constitutes an outlier precedes decisions concerning the identification and analytical judgments regarding their handling, and is often not without dialogue (e.g., Gnanadesikan & Kettenring, 1972). While it could be argued that outliers are distinct from measurement error, or are a special kind of measurement error, their presence in observational studies can hardly be debated. Indeed, one might argue that their presence -- and thus, influence -- is generally greater in observational studies because of lack of researcher control. For

purposes of this study, outlier contamination is viewed as a phenomenon that is not unlike measurement error in fundamental ways and the strategies for the conduct of this research were based on that central idea.

The purpose of the present study was to examine several regression methods within the framework of Weighted Structural Regression (WSR, Pruzek & Lepak, 1992), comparing both their regression weight stability and score estimation accuracy in the presence of outlier contamination. The specific regression methods to be compared are: Ordinary Least Squares, WSR Ridge Regression, Minimum Risk Regression, Minimum Risk*2, GFI, and WSR Reduced Rank Regression (cf. Pruzek & Lepak, 1992 for the latter five methods).

Method

Data Set Selection and Creation

This study used three "population" covariance matrices, in the form of correlation matrices, that have been drawn from applied behavioral science literature as the basis for generating samples -- some of which were "contaminated." All simulated raw data were, in turn, used to compare selected regression methods. Specifically, the population systems used for this study are identified with the terms Hauser[1] (1973), Language (Fuller, 1987), and Pitprop (Jeffers, Chap. 6, p. 176, in Mardia, Kent, & Bibby, 1980).

---

[1] This problem set was used by Rabinowitz (1990) and will, therefore, provide a linkage to recent work related to WSR as a class of methods.

Construction of the Contaminated Data

The contamination data generation process is begun first by constructing a matrix that is both multivariate normal and stochastically consistent with the chosen population system. This is done by performing a complete principal component factoring to obtain a coefficient matrix F_{ppc}, of order p x p, and premultiplying this by a matrix, of order n x p, constructed using a pseudo random normal generator [i.e., $x_{ij} \approx$ iid N(0,1)]. This results in a matrix, X_{sim}, that has the appropriate dimensionality (n x p) and is structurally consistent with the population system. Following the construction of X_{sim} a second matrix is constructed, X_{contam}. This matrix has an initial sample size considerably smaller than the population matrix (say only 10% of n, but with the appropriate p columns), is constructed using the same pseudo random normal generator, but is not constructed to be structurally consistent with the population system. This matrix is then premultiplied by a scalar (i.e., 3) to increase its variance beyond 1. Following this variance modification, the matrix has added to it the appropriate number of row, all entries being zero (0) to match the n x p dimensionality of X_{sim}. This matrix, X_{contam}, is then added to the X_{sim} matrix. The result is a predetermined number of rows (i.e., *k*) of X_{sim} having added to them some potentially extreme values from X_{contam}. The form of the contaminated data systems, $X_{sim\,|\,contam}$, is thus:

(1) $$X_{sim\,|\,contam} = X_{sim} + X_{contam}$$

where X_{contam} has dimensions n x p, with a predetermined number of contamination rows (k = n*%, % is the desired percentage of rows contaminated) that have a variance equal to 3. This contaminated data generation has been found to result in Kurtosis statistics (i.e., normal = 3) that are now consistently larger than normal.

Three different population models were used for each population: one normal distribution and two that were contaminated normals (see Method for details concerning their construction) with heavier than normal tails. Gleason (1993) states that contaminated normal distributions (CND) have several virtues, one of which is that they seem "intuitively plausible as a mechanism for inoculating an otherwise clean batch of data with the occasional outlier" (p. 327). While he dedicates the remainder of his paper to outlining that CNDs can be somewhat difficult to control, the workability and "intuitiveness" of this form of outlier contamination remains. Therefore, after systematic experimentation with the percentage of rows (i.e., k) to be contaminated it was decided to utilize two levels for k with each sample size (i.e., 60 & 120): 2 and 5 percent. These choices of k regularly resulted in kurtosis levels of 3.5 and 7.00, respectively (kurtosis is 3.00 for a normal). Therefore, in order to maintain data contamination levels that both simulated outliers, as evidenced by the increased kurtosis, while not overtly destroying the distributional characteristics of the data beyond a reasonable level (see Gleason, 1993), the contamination levels were set at 2 and 5 percent.

Bootstrap Method

In order to compare regression methods normal sampling procedures were used to generate 100 $X_{sim\,|\,contam}$ matrices for each population Σ. Efron and Tibshirani (1986) would describe this as a *contaminated* normal bootstrap procedure and this language will be used hereafter. Each of the 100 $X_{sim\,|\,contam}$, for each Σ, was constructed in the same manner for two different sample sizes -- 60 and 120. This method, which is essentially equivalent to Monte Carlo sampling, and sample sizes were also used by Rabinowitz (1990) and Datta (1992) and will, therefore, provide an additional linkage to extant WSR literature.

Regression Methods

Six different regression methods were examined using the aforementioned methodology and within the Weighted Structural Regression framework (Pruzek & Lepak, 1992, here after referred to as P & L). These six methods are: Ordinary Least Squares, WSR Reduced Rank Regression, Minimum Risk Regression, Minimum Risk*2, WSR Ridge Regression, and GFI. These six methods are briefly described in Appendix A (cf. P & L for more specific details).

Weighted Structural Regression as a Class of Methods

As described by P & L, Weight Structural Regression (WSR) uses a empirical Bayes covariance estimator taking the form:

(2) $$\hat{\Sigma}_{WSR-cfa} = w\,\hat{\Sigma}_{mf} + (1-w)\,\hat{\Sigma}_{no-bad} \quad .$$

Given the familiar OLS model for the calculation of regression weights,

$$\hat{\beta}_{ols} = \hat{\Sigma}_{xx}^{-1} \hat{\sigma}_{xy} \quad ,$$

and making the necessary substitution using the empirical Bayes matrix of equation (2), the general form of the WSR regression weight matrix $\hat{\beta}_{WSR-cfa}$ can be shown to take the form:

$$(3) \qquad \hat{\beta}_{WSR-cfa} = (w\hat{\Sigma}_{mf-xx} + (1-w)\hat{\Sigma}_{m-bsd-xx})^{-1}(w\hat{\sigma}_{mf-xy} + (1-w)\hat{\sigma}_{m-bsd-xy}) \quad .$$

(See P & L for details). From (3) it can be seen that if w is set to 1, then the regression weights are equivalent to OLS and if w is set to 0, the regression weights become a Reduced Rank solution. It is, therefore, by specifying w and selecting a particular common factor model (i.e., choice of m) that one can derive the desired set of regression methods. The adaptive forms of WSR -- Minimum Risk Ridge, Minimum Risk, Minimum Risk*2, and GFI -- utilize an empirical approach to determining w in which a squared error loss precedure (cf. Pruzek, 1993) is used to determine the relative goodness-of-fit of the cfa model, for the purpose of assigning the relative weights of the two population covariance estimators (model free & cfa model). That is, the better a cfa model "fits," the smaller w becomes, giving more weight to the cfa estimator of the convex sum. However, note that if $m = 0$, the result is not a cfa model, but a form of Minimum Risk Ridge Regression (cf. Appendix A). Equation (3) makes explicit how each set of regression weights is generated. It is from these

regression weights that the evaluative criteria of *Accuracy of Prediction* and *Stability of Beta Weights* data were obtained (see below).

It should be noted that the joint population criterion-predictor covariance matrices were rescaled using the uniqueness variance estimation procedure suggest by P & L. Additionally, the number of factors used for each population covariance matrix (correlation matrix) was specified according to an information-theoretic criterion (Bozdogan, 1990), but details of this procedure are omitted here (cf. Pruzek, 1993). In this presentation results are given for only a single value of m, but a larger simulation study involves two other values of m for each combination of population system, contamination level, and sample size (Rule, In progress).

Specific Forms for the Methods

Ordinary Least Squares: If $w = 1$ in (1), thus giving model free estimator component full weight within the covariance empirical Bayes, it can be shown that (3) reduces to:

$$\hat{\beta}_{WSR-cfa} = \hat{\beta}_{OLS} \quad .$$

WSR Reduced Rank: By setting $w = 0$ in (1), a WSR Reduced Rank version (for any particular value of m) of (3) is obtained having the form:

$$\hat{\beta}_{WSR-cfa} = \hat{\beta}_{RedR} \quad .$$

Minimum Risk (MinR): Minimum Risk Regression $\hat{\Sigma}_{WSR-cfa-MinR}$ (P & L) is a regression model that focuses upon the weighting function of the WSR formula (i.e., w), providing an adaptive procedure for its determination. Using procedures outlined within their paper, P & L have

developed a means of using information from the sample (size = n) to derive the statistic gamma, γ_m (cf. p. 104). In brief, gamma is computed in such a fashion as to adjust w in relation to how well the common factor structure fits the extant data. That is, the better the common factor structure (as defined by the selection of m) is supported by the data, the larger gamma becomes and the more emphasis will be placed upon the model based estimator within the empirical Bayes (i.e., $\hat{\Sigma}_{m-bsd}$). See Appendix A and P & L (cf. p. 104) for more specific details.

Minimum Risk 2 (MinR*2): Within the framework of Minimum Risk there may be instances when the researcher wishes to weight the structural model more heavily than what the data system would normally suggest during the calculation of gamma (i.e., γ_m) in the determination of w_{MinR}. In such instances the relative emphasis given to the structural model can be increased by pre multiply γ_m by a scalar -- 2 in the present case. The result is to down weight the model free estimator, $\hat{\Sigma}_{mf}$, and increase the influence of the model based estimator $\hat{\Sigma}_{m-bsd}$ in the empirical Bayes $\hat{\Sigma}_{WSR-cfa-MinR2}$.

WSR Ridge: By setting the number of factors, m, to zero the matrix $\hat{\Sigma}_{m-bsd}$ becomes *diagonal* and will be denoted $\hat{D}_{m-bsd-Rdg}$. The contribution of $\hat{D}_{m-bsd-Rdg}$ will be determined by using the same weighting function described for $\hat{\Sigma}_{WSR-cfa-MinR}$.

GFI: Called "Goodness of Fit Index" (GFI) by P & L, this method, like Minimum Risk and Minimum Risk 2, adjusts the weighting function within (2). Briefly, GFI is an index based on Jöreskog & Sorbom (1986)

adjusted goodness of fit measure as detailed by P & L in their paper (see Appendix A for more details).

Analysis

Stability of Beta Weights

The assessment of beta weight stability was assessed for each combination of Σ, n, k, and w (i.e., as above). Following each analysis cycle the sample betas (i.e., $\hat{\beta}_j$) for each regression method that are stored were used for the calculation of the average mean squared errors of sample (MSE_{beta}). The computation of MSE_{beta} took the following form:

$$MSE_{beta} = Avg(\hat{\beta}_j - \beta_{pop})^2$$

where $\hat{\beta}_j$ are the betas which result from each analysis cycle (i.e., using $X_{sim\,|\,contam}$) and β_{pop} are the OLS betas produced directly from Σ.

Additionally, the average of the estimable parts of the predictive mean squared errors ($EPMSE_{beta}$) was also calculated (P & L). Computation of $EPMSE_{beta}$ took the following form:

$$EPMSE_{beta} = Avg(\hat{\beta}_j - \beta_{ols})' R_{pop} (\hat{\beta}_j - \beta_{ols})$$

where $\hat{\beta}_j$ is the weight for each bootstrap sample, $\hat{\beta}_{ols}$ represents the population system regression weights, and R_{pop} represents the population correlation matrix. According to P & L (1992, p. 113) "The logic of the PMSE is that repeated use of $\hat{y}$s from samples in place of $\hat{y}$ from the population to predict y will result in a mean squared prediction error, per prediction, of σ^2 + E(EPMSE) where the latter term is the expected value

of EPMSE and σ^2 represents the residual population variance (Browne, 1975)."

Furthermore, a measure of bias was obtained for the different regression methods by averaging the stored $\hat{\beta}_j$ values across the contaminated normal bootstraps and comparing these values to the population β_{pop} with $(\text{Avg}(\hat{\beta}_j) - \beta_{pop})$.

Accuracy of Prediction

Following each analysis cycle -- the creation of $X_{sim\,|\,contam}$, the contaminated normal bootstrap sample, and the passing of the different regression methods over each simulated data set -- regression results were stored and summarized for each combination of Σ, n, k, and w.

As described in (1), the construction of $X_{sim\,|\,contam}$ has as one of its components an uncontaminated data system, X_{sim}. It was therefore possible to use the original simulated data before contamination as a target for assessing prediction accuracy of $\hat{y}$'s generated by each method.

Following the generation of each regression method's sample regression weights (i.e., $\hat{\beta}_j$), the weights were used to generate the predicted values of the criterion variable (i.e., **y**). Indeed, the $\hat{y}$ statistics were always based on independent samples of the same size, always applying the weights to simulated data before contamination. The mean squared errors of prediction, $MSE_{pred} = AVG(y - \hat{y})^2$, served as the criteria by which score accuracy was judged. This statistic was computed for every one of the 100 contaminated normal bootstrap samples by using a 'bootstrap cross-validation' procedure.

Results

The analysis of this study resulted in four sets of 100 bootstrap samples for each of the population systems; i.e., two sets corresponding to the two varying levels of outlier contamination for each of the two sample sizes utilized for the three data sets -- twelve sets in all. Although the procedures utilized for this study yielded information for all variables simultaneously, information is presented for only the criterion variable in order to simplify this presentation. In the case of the Hauser (Hauser, 1973) and Pitprop (Jeffers, Chap. 6, p. 176, in Mardia, Kent, & Bibby, 1980) the criterion variable was selected by the original author. In the case of the Language (Fuller, 1987) data set the first variable was arbitrarily chosen as the criterion. Results will be presented for each of the three populations in turn.

Hauser Data Set

Preliminary analysis (using an information-theoretic analysis, Bozdogan, 1990) showed that three common factors would be most appropriate for this population. (See Appendix B, Table I, for the Product-Moment Correlation Matrix and the Eigen values.) This value of three was used for all results for this population. For each combination of n (60 & 120), m (0 in the case of Minimum Risk Ridge), and k (2 & 5 percent) six different values of w and m were used to generate the desired regression methods (see Appendix A).

Stability of Beta Weights -- Hauser Data Set

Table 1 summarizes the overall MSE and EPMSE results for the Hauser population.

Table 1
Ratios of Average Mean Squared Errors of Beta -- Hauser Data Set (m = 3)
(First Variable was the criterion, all others predictors)

	Mean Squared Errors[a]				Pred. Mean Squared Errors[b]			
	k = 2%		k= 5%		k = 2%		k= 5%	
Method[c]	n=60	n=120	n=60	n=120	n=60	n=120	n=60	n=120
OLS	.29	.22	.51	.33	.15	.10	.27	.16
MR Ridge	1.76	1.31	1.80	1.36	1.52	1.20	1.56	1.23
MinR	2.00	1.49	1.68	1.39	1.64	1.29	1.47	1.24
MinR*2	2.29	1.67	1.98	1.59	**1.73**	**1.33**	**1.60**	**1.31**
GFI	**2.38**	**1.74**	**2.20**	**1.81**	1.67	1.21	**1.60**	1.22
RedR	2.12	1.61	2.10	1.62	1.41	1.06	1.28	.95

[a] The first row is the actual OLS Mean Squared Error (MSE). Subsequent rows give the ratios of the overall OLS MSE_{beta} to the MSE_{beta} of the given method. Ratios greater than 1.00 indicate advantages of the alternative methods.

[b] The first row is the actual OLS Predicted Mean Squared Error (EPMSE) followed by the EPMSE ratios for the other methods. $PMSE_{beta}$ ratios are calculated in a manner identical to that of the MSE_{beta} ratios.

[c] OLS -- Ordinary Least Squares, MR Ridge -- Minimum Risk, scale free, Ridge Regression, MinR (m = 0) -- Minimum Risk, MinR*2 -- Minimum Risk *2, GFI -- Goodness of Fit, and RedR -- Reduced Rank (w = 0).

As can be seen in Table 1, in all cases the non-OLS regression methods yielded greater regression weight stability than least squares regression -- the bold faced result indicates the "best" method for each condition. The greatest advantage for the empirical Bayes methods appeared with the smaller sample size (i.e., n =60). The three WSR

methods that weighted structural information most heavily (i.e., MinR*2, GFI, & RedR) consistently out-performed the methods using less structural information (i.e., WSR Ridge & MinR). In all cases the OLS method produced the highest overall MSE's and EPMSE's. Interestingly, only the Minimum Risk Ridge method showed a consistent improvement in the higher contamination conditions. Additionally, the average (across the four n and *k* conditions -- 100 samples each) *w* for GFI was .21 and for MinR*2 the average *w* was .48. (Smaller *w*'s signifying a greater reliance on the CFA model in the empirical Bayes convex sum).

Regression weight bias was obtained for the different regression methods by averaging the squared difference for the $(\hat{\beta}_j - \beta_{pop})$ values across the 100 contaminated normal samples. As one would expect, with an increasing reliance upon structural information (i.e., smaller *w*'s) a corresponding increase in regression weight bias results, with OLS displaying the least average squared bias and Reduced Rank displaying the greatest. This trend remained constant across both sample sizes and contamination conditions. (See Appendix C, Table I, for details of this analysis.)

Accuracy of Prediction -- Hauser Data Set

Mean squared errors of prediction, $MSE_{pred} = AVG(y - \hat{y})^2$, are summarized in Table 2.

As can be seen in Table 2, with one exception (Reduced Rank, n = 120, *k* = 5%) all non-OLS regression methods examined resulted in an increased prediction accuracy. Further, the prediction accuracy again

showed the expected pattern. That is, the advantage of regressions that utilized structural information in an adaptive form, via WSR (i.e., MinR, MinR*2, & GFI), was greatest when sample sizes were small (i.e., n =60) and outlier contamination was relatively high (i.e., 5%). The Minimum Risk* 2 method tended to work best across both sample sizes and contamination levels, but all of the non-OLS methods worked relatively well with regards to the criterion.

Table 2

Ratios of Average Squared Errors of Prediction Across Regression Methods -- Hauser Data Set (m = 3)[a]

	k = 2%		k= 5%	
Method[b]	n=60	n=120	n=60	n=120
OLS	.07	.07	.09	.07
MR Ridge	1.08	1.03	**1.12**	1.04
MinR	1.09	**1.04**	1.10	1.05
MinR*2	**1.10**	**1.04**	**1.12**	**1.06**
GFI	**1.10**	1.03	**1.12**	1.04
RedR	1.08	1.02	1.07	.98

[a] The first row is the actual OLS Mean Squared Error of Prediction (MSE_{pred}) multiplied by 10 to reduce the number of decimal places. Subsequent rows give the ratios of the overall OLS MSE_{pred} to the MSE_{beta} of the given method. Ratios greater than 1.00 indicate advantages of the alternative methods.

[b] OLS -- Ordinary Least Squares, MR Ridge -- Minimum Risk, scale free, Ridge Regression, MinR (m = 0) -- Minimum Risk, MinR*2 -- Minimum Risk *2, GFI -- Goodness of Fit, and RedR -- Reduced Rank (w = 0)

Language Data Set

As with the Hauser (1973) population system, an information-theoretic analysis indicated that three common factors would be most appropriate for the Language population system (Fuller, 1987). (See Appendix B, Table II, for the Product-Moment Correlation Matrix and the Eigen values.) Again, for each combination of n (60 & 120), *m* (0 in the case of Minimum Risk Ridge), and *k* (2 & 5 percent) six different values of *w* and m were used to generate the desired regression methods (see Appendix A).

Stability of Beta Weights -- Language Data Set

Table 3 summarizes the overall MSE and EPMSE results for the Language population. As with the Hauser data set results (see Table 1), in all cases the non-OLS regression methods yielded greater regression weight stability than least squares regression. The one exception to this occurred with the $EPMSE_{beta}$ for Minimum Risk Reduced Rank regression with n = 120 and *k* = 2%. The predicted pattern again emerged with greatest advantage for non-OLS regressions using structural information appear when sample size is small (i.e., n =60). The WSR methods of Minimum Risk*2 and GFI consistently out-performed the methods using less structural information (i.e., larger *w*'s) or none (i.e., OLS) in all conditions, there was a drop of regression weight stability ratios in the higher contamination conditions.

Table 3
Ratios of Average Mean Squared Errors of Beta - Language Data Set (m = 3)

	Mean Squared Errors[a]				Pred. Mean Squared Errors[b]			
	k = 2%		k= 5%		k = 2%		k= 5%	
Method[c]	n=60	n=120	n=60	n=120	n=60	n=120	n=60	n=120
OLS	.22	.18	.48	.40	.08	.07	.19	.15
MR Ridge	1.37	1.15	**1.33**	1.19	1.27	1.11	**1.25**	1.15
MinR	1.39	1.21	1.21	1.17	1.28	1.14	1.15	1.12
MinR*2	1.48	**1.28**	1.29	1.25	**1.34**	**1.17**	1.18	1.18
GFI	**1.50**	1.23	**1.33**	**1.36**	1.32	1.07	1.18	**1.20**
RedR	1.40	1.11	1.29	1.32	1.21	.95	1.08	1.14

[a] Refer to Table 1 for legend. [b] Refer to Table 1. [c] Refer to Table 1.

However, the Minimum Risk Ridge Regression showed a consistent improvement in the higher contamination conditions and resulted in the "best" $EPMSE_{beta}$ when n = 60 and k = 5%. The average (across the four n and k conditions -- 100 samples each) w for GFI was .19 and for MinR*2 the average w was .81. (Smaller w's signifying a greater reliance on the CFA model in the empirical Bayes convex sum).

Regression weight bias was again obtained for the different regression methods by averaging the squared difference for the $(\hat{\beta}_j - \beta_{pop})$ values across the 100 contaminated normal samples. Again, as one would expect (and found with the Hauser data set as well), with an increasing reliance upon structural information (i.e., smaller w's) a corresponding increase in regression weight bias results, with OLS displaying the least average squared bias and Reduced Rank displaying the greatest. Again,

this trend remained constant across both sample sizes and contamination conditions. (See Appendix C, Table II, for details of this analysis.)

Accuracy of Prediction -- Language Data Set

Mean squared errors of prediction, $MSE_{pred} = AVG(y - \hat{y})^2$, are summarized in Table 4.

As can be seen in Table 4, with one exception (Reduced Rank, n = 120, k = 2%) all non-OLS regression methods examined resulted in an increased prediction accuracy. Further, the prediction accuracy again showed the expected pattern, but, unlike the Hauser data set (see Table 2), remained relatively stable across sample size and contamination levels. Interestingly, the relative advantages of the non-OLS methods, as a whole, do not appear to result in as large gains in prediction accuracy (e.g., a maximum of 7%), as in the Hauser data (see Table 2) where the gains were often greater than 10% in magnitude.

(Continued on Next Page)

Table 4

Ratios of Average Squared Errors of Prediction Across Regression Methods -- Language Data Set (m = 3)[a]

Method[b]	k = 2% n=60	k = 2% n=120	k= 5% n=60	k= 5% n=120
OLS	.04	.04	.05	.05
MR Ridge	1.05	1.02	**1.07**	1.05
MinR	1.05	1.03	1.05	1.04
MinR*2	**1.06**	**1.04**	1.06	1.05
GFI	**1.06**	1.03	1.06	**1.07**
RedR	1.05	1.00	1.03	1.05

[a] Refer to Table 2 for legend. [b] Refer to Table 2.

Pitprop Data Set

Both of the two previous data sets were selected, in part, because they represent data typically found within the behavioral sciences. The final data set, Pitprop (Jeffers, Chap. 6, p. 176, in Mardia, Kent, & Bibby, 1980), relates to the pit prop poles used in mine shafts and was selected as a challenge to the factor-based WSR methods, particularly since these data do not appear to be characterized as having non-ignorable measurement error.

For this population the information-theoretic criterion suggested six (or possibly 5) factors. All results below were obtained with m = 6. (See Appendix B, Table III, for the Product-Moment Correlation Matrix and the Eigen values.)

Stability of Beta Weights -- Pitprop Data Set

Table 5 gives results for the Pitprop population. While the advantage for the adaptive WSR methods is still clearly evident and, at times, extremely large (i.e., MinR Ridge at k = 5% & n = 60), the methods that show the greatest gains tended to be those which relied less heavily upon the structural information (e.g., MinR Ridge & MinR). This result contrasts with the results from the two previous populations which generally favored WSR methods that placed more emphasis on the CFA model (see Tables 1 & 3).

For the MSE and PMSE criteria the Minimum Risk method emerged as the "best" one here, clearly out performing the other methods in the low contamination level conditions, particularly for small n. In the high contamination conditions (k = 5%) the MR Ridge method worked, best (MSE) with a 5% contamination level, but the Minimum Risk and Minimum Risk*2 methods were most effective in terms of the PMSE criteria for both levels and contaminations. The average (across the four n and k conditions -- 100 samples each) w for GFI was .41 and for MinR*2 the average w was .62. (Smaller w's signifying a greater reliance on the CFA model in the empirical Bayes convex sum).

Regression weight bias was again obtained for the different regression methods across the 100 contaminated normal samples. Unlike the bias results from the behavioral science populations, the bias results from the Pitprop data sets displayed a pattern that more closely resembles

the results of the regression weight stability results (i.e., Table 5 and Appendix C, Table III).

Table 5
Ratios of Average Mean Squared Errors of Beta - Pitprop Data Set (m = 6)

	Mean Squared Errors[a]				Pred. Mean Squared Errors[b]			
	k = 2%		k= 5%		k = 2%		k= 5%	
Method[c]	n=60	n=120	n=60	n=120	n=60	n=120	n=60	n=120
OLS	.91	.67	1.61	1.44	.11	.07	.16	.16
MR Ridge	1.00	.91	**1.33**	1.22	1.07	.94	1.05	1.05
MinR	**1.21**	**1.09**	1.23	1.15	**1.19**	**1.07**	**1.16**	1.09
MinR*2	1.18	1.06	1.27	1.21	**1.19**	1.04	**1.16**	**1.10**
GFI	1.13	.89	1.27	**1.26**	1.13	0.84	1.13	1.07
RedR	.88	.69	1.19	1.23	.85	0.59	.88	.87

[a] Refer to Table 1 for legend. [b] Refer to Table 1. [c] Refer to Table 1.

That is, the more able a regression method was to re-capture the population's predictor beta weight information, the less the resulting bias. For example, Minimum Risk displayed the lowest MSE_{beta} at the 2% contamination level with a sample size of 60 (MSE_{beta}= 1.21) and returned the lowest Average Squared Regression Weight Bias for that condition (i.e., 2.43). Readers are invited to refer to the two tables (i.e., Table 5 and Appendix C, Table III) for more details.

Accuracy of Prediction -- Pitprop Data Set

The prediction accuracy results are summarized in Table 6.

Table 6

Ratios of Average Squared Errors of Prediction Across Regression Methods -- Pitprop Data Set (m = 6)[a]

	k = 2%		k= 5%	
Method[b]	n=60	n=120	n=60	n=120
OLS	.04	.03	.04	.04
Ridge	1.03	.98	1.02	1.02
MinR	**1.05**	**1.01**	**1.05**	**1.03**
MinR*2	**1.05**	1.00	**1.05**	**1.03**
GFI	1.04	.96	1.04	1.02
RedR	.96	.86	.93	.95

[a] Refer to Table 2 for legend. [b] Refer to Table 2.

Like the regression weight results in Table 5 the results for the Pitprop problem show smaller gains. In no situation was the systematic advantage of any WSR method more that 5% for this criterion, and for the larger sample size with a low level of contamination the best WSR method (MinR) yielded a systemic advantage of only 1%. The Reduced Rank regression method consistently performed less well than the target OLS method with respect to this criterion and, in one case, so did the WSR form of Ridge regression (k = 2%, n = 120).

Discussion

The foregoing results help quantify the merits and possible demerits of the new WSR adaptive regression methods in three different populations relative to the ordinary least squares method on the basis of three criteria. The results of this study provide evidence supporting the

notion of the increased efficacy of the adaptive forms of WSR in small sample applications where the presence of outlier contamination exists. While showing that the improvement over conventional least squares is not always substantial, it is notable that adaptive forms of WSR based on the concept of empirical Bayes covariance estimation can provide consistent and sometimes substantial improvement over conventional methods. It is also important to recognize that the nature of the gains or losses must attend to the specific evaluative criterion as well as the nature of the parent population. Additionally, one virtue of these methods lies in the possibility of increased accuracy in the area of outlier detection. That is, if one is able to more accurately recover both β_{pop} and $\hat{y}$, then these may provide a basis for a re-analysis of data for the express purpose of outlier detection.

Although not reported in this presentation, ongoing examination of these WSR methods suggests that the choice of *m*, the number of factors, is not as critical as one might expect (Rule, in progress). That is, although the results do tend to fluctuate when utilizing different values for *m* (say ± 1), the pattern of results reported here are generally replicated. These findings suggest that researchers with less than omniscient knowledge concerning a population system's cfa structure can expect to obtain gains similar to those reported within this presentation.

In conclusion, the present study provides additional evidence to support the exploration of the WSR class of methods in the realm of small sample problems within social science population systems or systems for

which the use of the underlying empirical Bayes covariance estimation procedure's use of a cfa model is justifiable.

References

Andrews, D. F., & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society, Ser. B, 40*(1), 85-93.

Bozdogan, H. (1990). On the information based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics - Theory Methods, 19*, 221-278.

Darlington, R. B. (1990). *Regression and Linear Models*. New York: McGraw-Hill.

Datta, P. (1992). *A simulation study of cononical/inter-battery methods based on convex sum procedures*. Unpublished doctoral dissertation, University at Albany, State University of New York.

Dillon, W. R., & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. New York: John Wiley & Sons.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54-77.

Fuller, W. A. (1987). *Measurement error models*. John Wiley & Sons: New York.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics, 28*, 81-124.

Gleason, J. R. (1993). Understanding elongation: The scale contaminated normal family. *Journal of the American Statistical Association, 88*(421), 327-337.

Hadi, A. S. (1985). K-clustering as a detection tool for influential subsets in regression (with discussion). *Technometrics, 27*, 323-325.

Hauser, R. M. (1973). Disaggregrating a model of educational attainment. In Goldberger, A. S., & Duncan, O. D., *Structural Equation Models in the Social Sciences*, 255-284. New York: Seminar Press.

Jöreskog, K. G., & Sorbom, D. (1986). *Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares (4th edition)*. Mooresville, IN: Scientific Software.

Mardia, K. V., Kent, J. R., & Bibby, J. M. (1980). Multivariate analysis. New York: Academic Press.

Rabinowitz, S. N. (1990). *A Simulation Study of a Class of Random Variable Linear Regression Methods*. Unpublished doctoral dissertation, Department of Educational Psychology and Statistic, State University of New York at Albany.

Pedhazur, E. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction* (2nd ed.). Montreal: Holt, Rinehart and Winston.

Pruzek, R. M. (1993). High dimensional covariance estimation: Avoiding 'the curse of dimensionality'. In H. Bozdogan (Ed.), *Multivariate Statistical Modeling, Vol. 2, Proceedings of First US/Japan conference on the frontiers of statistical modeling: An informational approach.* Amsterdam: Kluwer Academic Press.

Pruzek, R. M., & Lepak, G. M. (1992). Weighted structural regression: A broad class of adaptive methods for improving linear prediction. *Multivariate Behavioral Research*, 27(1), 95-129.

Appendix A

The following, detailed within Pruzek & Lepak (1992), provides the basis for the WSR class of methods. Readers are directed to P & L for further explanation and examples.

Prior to eigenanalysis, the rescaling of the joint predictor-criterion correlation matrix takes the form:

$$\hat{S}^{*-1} R_{*} \hat{S}^{*-1} = Q D_{\lambda}^{2} Q'$$

where $\hat{S}^{*-1}$ is the inverse of a diagonal matrix whose non-zero entries take a form consistent with Muirhead's (1985, cited in P & L) linear estimator to correct estimates of a function of *smc*, rescaling R* in the metric of the compliments of the *smc*'s. The result of this rescaling R^{*} then provides the starting point for eigenanalysis.

Empirical Bayes covariance estimation takes the general form:

(1) $$\hat{\Sigma}_{WSR-cfa} = w \hat{\Sigma}_{mf} + (1-w) \hat{\Sigma}_{m-bsd} \ .$$

where w is a scalar weighting function $0 \leq w \leq 1$, $\hat{\Sigma}_{mf}$ is the model free covariance estimator, and $\hat{\Sigma}_{m-bsd}$ is the cfa model estimator of $\hat{\Sigma}_{mf}$, and $\hat{\Sigma}_{WSR-cfa}$ is the WSR empirical Bayes estimator of Σ_{pop}.

If $w = 1$ no information from a cfa model is included in the empirical Bayes covariance estimator and ordinary least squares regression results. If $w = 0$ only information from the cfa model is used and Reduced Rank regression results.

Four WSR adaptive forms (MR Ridge, MinRisk, MinRisk*2, and GFI) use a quadratic loss approach to specifying the value of w. In the first two of these adaptive methods (MR Ridge and MinRisk) $w_m = n/(n+\hat{\gamma}_m)$ where $\hat{\gamma}_m$ is a function of m, p, $\hat{r}_m$ (ratio of trace function). In MinRisk*2 $w_m = n/(n+2\gamma_m)$ and for GFI w is based on a wholly different mechansm based on the fit of $\hat{\Sigma}_{m-bsd}$ to $\hat{\Sigma}_{mf}$ (cf. P & L, 1992).

Using (1), setting $m = 0$, and substituting w_m for w a form of Minimum Risk Ridge Regression empirical Bayes approach in determining the relative contribution of $\hat{D}_{m-bsd-Rdg}$ to the resultant $\hat{\Sigma}_{WSR-Rdg}$, which will take the ridge form:

$$\hat{\Sigma}_{WSR-Rdg} = w_m \hat{\Sigma}_{mf} + (1-w_m)\hat{D}_{m-bsd-Rdg} \quad .$$

Using (1), with some number of factors (cfa model), m, in the estimator yeilds Minimum Risk Regression. Premultiplying $\hat{\gamma}_m$ by a scalar of 2 results in Minimum Risk*2. The GFI method is obtained by setting:

$$1-w = \text{GFI} = 1-(d_1/d_2)(\text{tr}(\hat{\Sigma}_{m-bsd}^{-1}\hat{\Sigma}_{mf}-I)^2 / \text{tr}(\hat{\Sigma}_{m-bsd}^{-1}\hat{\Sigma}_{mf})^2)$$

where (d_1/d_2), a term that corrects for the degrees of freedom, is equal to $p(p+1)/((p-m)^2+p-m+2)$ (cf. Jöreskog & Sorbom, 1986).

Appendix B

Table I
Population System -- Hauser Data

Matrix of Product-Moment Correlations

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00											
2	.51	1.00										
3	.49	.32	1.00									
4	.39	.29	.52	1.00								
5	.24	.23	.21	.20	1.00							
6	.15	.15	.13	.12	.59	1.00						
7	.16	.14	.14	.15	.35	.44	1.00					
8	.30	.27	.29	.29	.37	.34	.42	1.00				
9	.28	.26	.28	.29	.32	.32	.33	.42	1.00			
10	.31	.27	.30	.30	.43	.47	.44	.54	.50	1.00		
11	.29	.25	.30	.29	.46	.47	.41	.51	.47	.77	1.00	
12	.34	.29	.33	.32	.48	.55	.41	.49	.49	.66	.59	1.00

Eigen values

1	2	3	4	5	6	7	8	9	10	11	12
5.02	1.63	.86	.77	.68	.59	.53	.51	.43	.42	.34	.22

Table II
Population System -- Language Data

Matrix of Product-Moment Correlations

	1	2	3	4	5	6	7	8
1	1.00							
2	.80	1.00						
3	.49	.49	1.00					
4	.58	.58	.53	1.00				
5	.63	.72	.56	.57	1.00			
6	.53	.55	.62	.54	.58	1.00		
7	.51	.50	.67	.54	.54	.70	1.000	
8	.47	.45	.46	.69	.57	.45	.47	1.000

Eigen values

1	2	3	4	5	6	7	8
4.95	.82	.73	.40	.37	.29	.26	.17

Table III
Population System -- Pitprop Data

Matrix of Product-Moment Correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00													
2	.42	1.00												
3	.33	.95	1.00											
4	.73	.36	.30	1.00										
5	.54	.34	.28	.88	1.00									
6	.25	.13	.12	.15	-0.22	1.00								
7	-0.12	.31	.29	.15	.38	-0.36	1.00							
8	-0.11	.50	.50	-0.03	.17	-0.30	.81	1.00						
9	.25	.42	.42	-0.05	-0.06	-0.00	.09	.37	1.00					
10	.24	.59	.65	.13	.14	.04	.21	.47	.48	1.00				
11	.10	.56	.57	-0.08	-0.01	-0.04	.27	.68	.56	.53	1.00			
12	.06	.08	.08	.16	.10	.09	-0.04	-0.11	.06	.09	-0.32	1.00		
12	.12	-0.02	-0.04	.22	.17	.15	.03	-0.23	-0.36	-0.13	-0.37	.03	1.00	
14	-0.15	-0.13	-0.14	-0.13	-0.02	-0.21	.33	.42	.20	.09	.29	-0.01	-0.184	1.00

Eigen values

1	2	3	4	5	6	7	8	9	10	11	12	13	14
4.33	2.80	1.90	1.13	1.08	0.84	0.60	0.45	0.35	0.20	0.19	0.05	0.04	0.03

Appendix C

Table I

Hauser Data Squared Regression Weight Bias Values (times 10)

Row Sum of Predictor Variables*

	k = 2%		*k*= 5%	
Method**	n=60	n=120	n=60	n=120
OLS	.04	.09	.05	.05
MR Ridge	.07	.10	.16	.11
MinR	.11	.14	.19	.14
MinR*2	.18	.20	.29	.21
GFI	.27	.35	.43	.46
RedR	.43	.45	.75	.72

* This index is calculated by taking each method's Average Squared Regression Weight Bias for each predictor variable, multiplying it by 10 -- in order to bring the values to within a place or two of the decimal, and taking the average across all predictor variables.

** OLS -- Ordinary Least Squares, MR Ridge -- WSR Ridge Regression, MinR -- Minimum Risk, MinR*2 -- Minimum Risk *2, GFI -- Goodness of Fit, and RedR -- Reduced Rank

Table II
Language Data Squared Regression Weight Bias Values (times 10)
Row Sum of Predictor Variables*

	k = 2%		*k*= 5%	
Method**	n=60	n=120	n=60	n=120
OLS	.03	.05	.31	.33
MR Ridge	.22	.15	.68	.53
MinR	.15	.14	.55	.50
MinR*2	.24	.21	.68	.62
GFI	.40	.49	.92	.98
RedR	.53	.62	1.19	1.16

* See Above ** See Above

Table III
Pit Prop Data Squared Regression Weight Bias Values (times 10)
Row Sum of Predictor Variables*

	k = 2%		*k*= 5%	
Method**	n=60	n=120	n=60	n=120
OLS	.55	.63	1.55	4.12
MR Ridge	6.92	5.10	8.15	7.48
MinR	2.43	1.88	3.78	5.24
MinR*2	3.54	2.81	4.89	5.89
GFI	4.47	5.09	5.60	7.09
RedR	7.42	7.73	8.51	8.51

* See Above ** See Above